

OECD GUIDELINE FOR THE TESTING OF CHEMICALS

The Up-and-Down Procedure for Acute Oral Toxicity: Proposed Test Guideline

INTRODUCTION

1. OECD guidelines for the Testing of Chemicals are periodically reviewed in the light of scientific progress or changing assessment practices. The concept of the up-and-down testing approach was first described by Dixon and Mood (1)(2)(3)(4). In 1985, Bruce proposed to use an up-and-down procedure (UDP) for the determination of acute toxicity of chemicals (5). There exist several variations of the up-and-down experimental design for estimating an LD50. This guideline is based on the procedure of Bruce as adopted by ASTM in 1987 (6) and revised in 1990. A study comparing the results obtained with the UDP, the conventional LD50 test and the Fixed Dose Procedure (FDP, Guideline 420) was published in 1995 (7). Since the early papers of Dixon and Mood, papers have continued to appear in the biometrical and applied literature, examining the best conditions for use of the approach (8)(9)(10)(11). Based on the recommendations of several expert meetings in 1999, an additional revision was considered timely because: i) international agreement had been reached on harmonised LD50 cut-off values for the classification of chemical substances, ii) testing in one sex (usually females) is generally considered sufficient, and iii) revision was being undertaken concurrently for two other alternatives to the conventional acute oral toxicity test, described in Test Guideline 401.

2. This test procedure is of value in minimizing the number of animals required to estimate the acute oral toxicity of a chemical as indicated by an estimated LD50, given knowledge before testing of the approximate LD50 and slope. In addition to the observation of mortality, the test allows the observation of signs of toxicity. A supplemental procedure also allows estimation of the slope of the dose response curve.

3. Definitions of some terms are in Appendix I.

INITIAL CONSIDERATIONS

4. All available information on the test substance should be considered by the testing laboratory prior to conducting the study. Such information will include the identity and chemical structure of the substance; its physical chemical properties; the results of any other *in vitro* or *in vivo* toxicity tests on the substance; toxicological data on structurally related substances; and the anticipated use(s) of the substance. This information is necessary to satisfy all concerned that the test is relevant for the protection of human health, and will help in the selection of an appropriate starting dose.

5. When designing a UDP test, if no information is available to make a preliminary estimate of the LD50 and/or the slope of the dose response curve, results of computer simulations have suggested that starting near 175 mg/kg and using half-log units (corresponding to a dose progression of 3.2) between doses will produce the best results. The half-log spacing balances a more efficient use of animals, while reducing bias in the prediction of the LD50 value. Coupled

with this concern, in order that any bias will not lead to under-classification, it is essential that initial dosing occur below the estimated LD50. However, for chemicals with large variability (i.e., shallow dose-response slopes), simulations indicate that bias can still be introduced in the lethality estimates and the LD50 has a large statistical error, similar to other acute toxicity methods. To correct for this, the single-sequence test as described herein includes a stopping rule not keyed to a fixed number of test observations but to properties of the estimate. Although the stopping rule is applied to all data, simulations have shown that it will make no essential difference in animal usage for the great majority of chemicals.

6. The UDP is easiest to apply to materials that produce death within one or two days. The method would not be practical to use when considerably delayed death (five days or more) can be expected.
7. Computers are used to facilitate animal-by-animal calculations that establish testing sequences and provide final estimates.
8. During the test, all animals obviously in pain or showing signs of severe distress should be humanely killed.
9. A limit test can be used efficiently to identify chemicals that are likely to have low toxicity.

PRINCIPLE OF THE PRIMARY (SINGLE ESTIMATE) TEST

10. For each run, animals are dosed, one at a time, at 48 hour intervals. The first animal receives a dose a step below the level of the best estimate of the LD50. If the animal survives, the dose for the next animal is increased to a factor of 3.2 times the original dose; if it dies, the dose for the next animal is decreased by a similar dose progression. (Note: 3.2 is the default factor. Paragraph 20 provides further guidance for choice of dose spacing factor.) Each animal should be observed carefully for 48 hours (unless the animal dies) before making a decision on whether and how much to dose the next animal. That decision is based on the survival pattern of all the animals up to that time. A combination of stopping criteria is used to keep the number of animals low while adjusting the dosing pattern to reduce the effect of a poor starting value (see paragraph 20). Dosing may be stopped when an estimate of LD50 is obtained which satisfies these criteria (see paragraphs 20 and 33). In typical cases for most applications, testing will be completed with only 4 animals after initial reversal in animal outcome. In any event, the test uses no more than 15 animals. The LD50 is calculated using the method of maximum likelihood (12)(13). A description of the maximum likelihood procedure is in paragraphs 31 and 32.

PRINCIPLE OF THE SUPPLEMENTAL TEST

11. When an estimation of slope is desired, the primary procedure serves as the starting point for a tailored testing and estimation routine. The supplemental procedure also provides a confidence interval for the LD50. A description of this supplemental procedure starts at paragraph 22 and the formula for this calculation is provided in paragraph 34. It is based on the

principle that multiple sequences with associated LD50s give an estimate of the standard error of the estimate of the LD50, which is related to the slope in a known way.

DESCRIPTION OF THE METHOD

Selection of animals species

12. The preferred rodent species is the rat although other rodent species may be used. In the normal procedure, female rats are used because literature surveys of conventional LD50 tests show that, although there is little difference of sensitivity between sexes, in those cases where differences were observed, females were in general more sensitive. When there is adequate information to infer that males are more sensitive, they should replace females in the test.

13. Healthy young adult animals should be employed. Littermates should be randomly assigned to treatment levels. The females should be nulliparous and non-pregnant. At the commencement of the study, the weight variation of the animals should be minimal and not exceed $\pm 20\%$ of the mean weight for each sex. The test animals should be characterised as to species, strain, source, sex, weight and/or age.

Housing and feeding conditions

14. The temperature in the experimental animal room should be 22°C ($\pm 3^\circ\text{C}$). Although the relative humidity should be at least 30 % and preferably not exceed 60 % other than during room cleaning, the aim should be 50-60 %. Lighting should be artificial, the sequence being 12 hours light and 12 hours dark. The animals are housed individually. Unlimited supply of conventional rodent laboratory diets and drinking water should be provided.

Preparation of animals

15. The animals are uniquely identified and kept in their cages for at least five days prior to dosing for acclimatization to the laboratory conditions. During acclimatization the animals should be observed for ill health. Animals demonstrating signs of spontaneous disease or abnormality prior to the start of the study are eliminated from the study.

Preparation of doses

16. When necessary, the test substance is dissolved or suspended in a suitable vehicle. It is recommended that, whenever possible, the use of an aqueous solution or suspension be considered first, followed by consideration of a solution or emulsion in oil (e.g. corn oil) and then by possible solution in other vehicles. For vehicles other than water, the toxicity of the vehicle must be known. In rodents, the volume should not normally exceed 1 mL/100 g body weight; however, in the case of aqueous solutions 2 mL/100 g body weight can be considered.

PROCEDURE

Primary testing using a single-sequence of dosing

17. For selecting the starting dose, all available information should be used, including information on structure-activity relationships. When the information suggests that mortality is unlikely, a limit test should be conducted (see paragraph 23). When there is no information on the substance to be tested, it is recommended that the starting dose of 175 mg/kg body weight be used (see Appendix II). This dose serves to reduce the level of pain and suffering by starting at a dose which in most cases will be sublethal. In addition, this dose reduces the chance that hazard of the chemical will be underestimated.

18. For each run, single animals are dosed in sequence usually at 48 h intervals. However, the time intervals between dosing should not be fixed rigidly and may be adjusted as appropriate (e.g., in case of delayed mortality). The first animal is dosed a step below the toxicologist's best estimate of the LD50. If no estimate of the chemical's lethality is available, dosing should be initiated at 175 mg/kg. If the animal survives, the second animal receives a higher dose. If the first animal dies or appears moribund, the second animal receives a lower dose (see paragraph 20 for size of dose spacing). Animals killed for humane reasons are considered in the same way as animals that died on test. Dosing should not normally exceed 2000 mg/kg body weight. However, when justified by specific regulatory needs, testing up to 5000 mg/kg body weight may be considered.

19. Moribund state is characterised by symptoms such as shallow, labored or irregular respiration, muscular weakness or tremors, absence of voluntary response to external stimuli, cyanosis and coma. Criteria for making the decision to humanely kill moribund and severely suffering animals are the subject of the separate OECD *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation*

20. The dose for each successive animal is adjusted up or down, depending on the outcome of the previous animal. At the outset, if feasible, a slope of the dose response should also be estimated based on all information available to the toxicologist including structure activity relationships. The dose progression factor should be chosen to be the antilog of 1/(the estimated slope of the dose response curve). When there is no information on the substance to be tested, a dose progression factor of 3.2 is used. Dosing continues depending on the outcomes of all the animals up to that time. In any event, if 15 animals have been tested, testing stops. Prior to that, the test is stopped based on the outcome pattern if:

- (1) the upper testing bound is reached and 3 consecutive animals survive at that bound or if the lower bound is reached and 3 consecutive animals die at that bound, or
- (2) the next animal to be tested would be the 7th and each surviving animal to this point has been followed by a death and vice versa (i.e., 5 reversals occur in 6 animals started), otherwise;
- (3) evaluation whether testing stops or continues is based on whether a certain stopping criterion is met: Starting following the fourth animal after the first reversal (which may be as early as the decision about the seventh animal), three measures of test progress are

compared via two ratios. If the first measure is at least two-and-one-half times both the other measures (i.e., both ratios are 2.5), testing is stopped. (see paragraph 33 and Appendix III). For a wide variety of combinations of LD50 and slopes as low as 2.5, the stopping rule will be satisfied with four to six additional animals, with fortuitously well-placed tests using even fewer. However, for chemicals with shallow dose-response slope (large variance), more animals may be needed. If animal tolerances to the chemical are expected to be highly variable (i.e., slopes are expected to be less than 3), consideration should be given to increasing the dose progression factor beyond the default 0.5 log dose (i.e., 3.2 progression factor) prior to starting the test.

21. When the stopping criteria have been attained after the initial reversal, the LD50 should be calculated using the method described in paragraphs 31 and 32.

Supplemental Test: Estimate an LD50 and Slope of the Dose Response Curve

22. Following the primary test, a supplemental test to estimate the slope of the dose-response curve can be implemented when necessary. This procedure uses multiple testing sequences similar to the primary test, with the exception that the sequences are intentionally begun well below the LD50 estimate from the primary test. These test sequences should be started at doses at least 10 times less than the LD50 estimate from the primary test, and not more than 32 times less. Testing continues in each sequence until the first animal dies. Doses within each sequence are increased by the standard 3.2 factor. The starting doses for each test sequence should be staggered, as described in Appendix II, paragraph 6. Upon completion of up to six of these supplemental test sequences, a standard probit analysis should be run on the entire collection of data, including the outcomes of the primary test. Good judgment will be required in cases where the primary test yields estimates of LD50 that are too close to the lower limit of doses tested. When this occurs, testing may be required to begin well above the LD50, where deaths are likely, and each sequence will terminate with the first survivor. If slope may be highly variable, an alternate procedure, using varying dose progression sizes, may be appropriate as shown in Appendix IV.

Limit test

23. Dosing should not normally exceed 2000 mg/kg body weight. However, when justified by specific regulatory needs, testing up to 5000 mg/kg body weight may be considered. One animal is dosed at the upper limit dose; if it survives, two more animals are dosed sequentially at the limit dose; if both animals survive, the test is stopped. If one or both of these two animals die, two animals are dosed sequentially at the limit dose until a total of three survivals or three deaths occurs. If three animals survive, the LD50 is estimated to be above the limit dose. If three animals die, the LD50 is estimated to be at or below the limit dose. If the first animal dies, a primary test should be run to determine the LD50 (see paragraph 11 of appendix II). As with any limit test protocol, the probability of correctly classifying a compound will decrease as the actual LD50 approaches the limit dose. The selection of a sequential test plan increases the statistical power and also has been made to intentionally bias the procedure towards rejection of the limit test for compounds with LD50s near the limit dose, i.e., to err on the side of safety.

Administration of doses

24. The test substance is administered in a single dose to the animals by gavage using a stomach tube or a suitable intubation cannula. The maximum volume of liquid that can be administered at one time depends on the size of the test animal. In rodents, the volume should not normally exceed 1 ml/100 g body weight; however, in the case of aqueous solutions 2 ml/100 g body weight can be considered. When a vehicle other than water is used, variability in test volume should be minimised by adjusting the concentration to ensure a constant volume at all dose levels. If administration in a single dose is not possible, the dose may be given in smaller fractions over a period not exceeding 24 hours.

25. Animals should be fasted prior to dosing (e.g., with the rat, food but not water should be withheld overnight; with the mouse, food but not water should be withheld for 3-4 hours). Following the period of fasting, the animals should be weighed and the test substance administered. The fasted body weight of each animal is determined and the dose is calculated according to the body weight. After the substance has been administered, food may be withheld for a further 3-4 hours in rats or 1-2 hours in mice. Where a dose is administered in fractions over a period of time, it may be necessary to provide the animals with food and water depending on the length of the period.

Observations

26. After dosing, animals are observed individually at least once during the first 30 minutes, periodically during the first 24 hours, with special attention given during the first 4 hours, and at least once daily thereafter. The animals should normally be observed for 14 days, except where animals need to be removed from the study and humanely killed for animal welfare reasons or are found dead. However, the duration of observation should not be fixed rigidly. It should be determined by the toxic reactions, time of onset and length of recovery period, and may thus be extended when considered necessary. The times at which signs of toxicity appear and disappear are important, especially if there is a tendency for toxic signs to be delayed (14). All observations are systematically recorded with individual records being maintained for each animal. Toxicology texts should be consulted for information on the types of clinical signs that might be observed.

27. Careful clinical observations should be made at least twice on the day of dosing, or more frequently when indicated by the response of the animals to the treatment, and at least once daily thereafter. Animals found in a moribund condition and animals showing severe pain and enduring signs of severe distress should be humanely killed. When animals are killed for humane reasons or found dead, the time of death should be recorded as precisely as possible. Additional observations will be necessary if the animals continue to display signs of toxicity. Observations should include changes in skin and fur, eyes and mucous membranes, and also respiratory, circulatory, autonomic and central nervous systems, and somatomotor activity and behaviour pattern. Attention should be directed to observations of tremors, convulsions, salivation, diarrhoea, lethargy, sleep and coma.

Body weight

28. Individual weights of animals should be determined shortly before the test substance is administered, at least weekly thereafter, at the time of death or at day 14 in the case of survival. Weight changes should be calculated and recorded.

Pathology

29. All animals, including those which die during the test or are killed for animal welfare reasons during the test and those that survive at day 14, are subjected to gross necropsy. The necropsy should entail a macroscopic inspection of the visceral organs. As deemed appropriate, microscopic analysis of target organs and clinical chemistry may be included to gain further information on the nature of the toxicity of the test material.

DATA AND REPORTING**Data**

30. Individual animal data should be provided. Additionally, all data should be summarised in tabular form, showing for each test concentration the number of animals used, the number of animals displaying signs of toxicity (Chan and Hayes, 14), the number of animals found dead during the test or killed for humane reasons, time of death of individual animals, a description and the time course of toxic effects and reversibility, and necropsy findings. A rationale for the starting dose and the dose progression and any data used to support this choice should be provided.

Calculation of LD50 for the primary test

31. The LD50 is calculated using the maximum likelihood method (12)(13), other than in exceptional cases given below. The following statistical details may be helpful in implementing the maximum likelihood calculations suggested (with an assumed *sigma*). All deaths, whether immediate or delayed or humane kills, are incorporated for the purpose of the maximum likelihood analysis. Following Dixon (4), the likelihood function is written as follows:

$$L = L_1 L_2 \dots L_n ,$$

where

L is the likelihood of the experimental outcome, given μ and σ , and n the total number of animals tested.

$L_i = 1 - F(Z_i)$ if the i^{th} animal survived, or
 $L_i = F(Z_i)$ if the i^{th} animal died,

where

F = cumulative standard normal distribution,

$Z_i = [\log(d_i) - \mu] / \sigma$

d_i = dose given to the i^{th} animal, and

σ = standard deviation in log units of dose (which is not the log standard deviation).

When identifying the maximum of the likelihood L to get an estimate of the true LD50, μ is set = log LD50, and automated calculations solve for it (see paragraph 32).

An estimate of σ of 0.5 is used unless a better generic or case-specific value is available.

(a) If testing stopped based on criterion (1) (i.e., a boundary dose was tested repeatedly), or if the upper bound dose ended testing, then the LD50 is reported to be above the upper bound; if the lower bound dose ended testing then the LD50 is reported to be below the lower bound dose. Classification is completed on this basis.

(b) If all the dead animals have higher doses than all the live animals or, vice versa, the LD50 is between the doses for the live and the dead animals, these observations give no further information on the exact value of the LD50. Still, a maximum likelihood LD50 estimate can be made provided there is a value for σ . Stopping criterion (2) in paragraph 20 describes one such circumstance.

(c) If the live and dead animals have only one dose in common and all the other dead animals have higher doses and all the other live animals lower doses, or vice versa, then the LD50 equals their common dose. If there is ever cause to repeat the test, testing should proceed with a smaller dose progression.

If none of the above situations occurs, then the LD50 is calculated using the maximum likelihood method.

32. Maximum likelihood calculation can be performed using either SAS (12)(e.g., PROC NLIN) or BMDP (13)(e.g., program AR) computer program packages as described in Appendix 1D in Reference 3. Other computer programs may also be used. Typical instructions for these packages are given in appendices to the ASTM Standard E 1163-87 (6). The σ used in the BASIC program in (6) will need to be edited to reflect the changes in this version of the OECD 425 Guideline. The program's output is an estimate of log(LD50) and its standard error.

33. The stopping criterion (3) in paragraph 20 is based on three measures of test progress, that are of the form of the likelihood in paragraph 31, with different values for μ , and comparisons are made after each animal tested after the sixth that does not already satisfy criterion (1) or (2). The equations for criterion (3) are provided in Appendix III. These comparisons are most readily performed in an automated manner and can be executed repeatedly, for instance, by a spreadsheet routine such as that also provided in Appendix III. If the criterion is met, testing stops and the LD50 can be calculated by the maximum likelihood method.

Calculation of LD50 and Slope Using Supplemental Procedure

34. A Supplemental Procedure is based on running three independent replicates of the Up-and-Down Procedure. Each replicate starts at least one log, but not more than 1.5 log, below the estimated LD50. Each run stops when the first animal dies. All data from these runs and the original Up-an-Down run are combined and an LD50 and slope are calculated using a standard probit method.

Report

35. The test report must include the following information:

Test substance:

- physical nature, purity and physicochemical properties (including isomerisation);
- identification data.

Vehicle (if appropriate):

- justification for choice of vehicle, if other than water.

Test animals:

- species/strain used;
- microbiological status of the animals, when known;
- number, age and sex of animals;
- rationale for use of males instead of females;
- source, housing conditions, diet, etc.;
- individual weights of animals at the start of the test, at day 7, and at day 14.

Test conditions:

- rationale for initial dose level selection, dose progression factor and for follow-up dose levels;
- details of test substance formulation;
- details of the administration of the test substance;
- details of food and water quality (including diet type/source, water source).

Results:

- body weight/body weight changes;
- tabulation of response data by sex (if both sexes are used) and dose level for each animal (i.e. animals showing signs of toxicity including nature, severity, duration of effects, and mortality);
- time course of onset of signs of toxicity and whether these were reversible for each animal;
- necropsy findings and any histopathological findings for each animal, if available;
- slope of the dose response curve (when determined);

- LD50 data;
- statistical treatment of results (description of computer routine used and spreadsheet tabulation of calculations)

Discussion and interpretation of results.

Conclusions.

LITERATURE

- (1) Dixon, W.J. and A.M. Mood. (1948). A Method for Obtaining and Analyzing Sensitivity Data. J. Amer. Statist. Assoc., 43, 109-126.
- (2) Dixon, W.J. (1965). The Up-and-Down Method for Small Samples. J. Amer. Statist. Assoc., 60, 967-978.
- (3) Dixon, W.J. (1991). Staircase Bioassay: The Up-and-Down Method. Neurosci. Biobehav. Rev., 15, 47-50.
- (4) Dixon, W.J. (1991) Design and Analysis of Quantal Dose-Response Experiments (with Emphasis on Staircase Designs). Dixon Statistical Associates, Los Angeles, CA, USA.
- (5) Bruce, R.D. (1985). An Up-and-Down Procedure for Acute Toxicity Testing. Fundam. Appl. Tox., 5, 151-157.
- (6) ASTM (1987). E 1163-87, Standard Test Method for Estimating Acute Oral Toxicity in Rats. American Society for Testing and Materials, Philadelphia, PA, USA.
- (7) Lipnick, R.L., J.A. Cotruvo, R.N. Hill, R.D. Bruce, K.A. Stitzel, A.P. Walker, I. Chu, M. Goddard, L. Segal, J.A. Springer, and R.C. Myers. (1995). Comparison of the Up-and-Down, Conventional LD₅₀ and Fixed Dose Acute Toxicity Procedures. Fd. Chem. Toxicol., 33, 223-231.
- (8) Choi, S.C. (1990). Interval estimation of the LD₅₀ based on an up-and-down experiment. Biometrics 46, 485-492.
- (9) Vågerö, M. and R. Sundberg. (1999). The distribution of the maximum likelihood estimator in up-and-down experiments for quantal dose-response data. J. Biopharmaceut. Statist., 9(3), 499-519.
- (10) Hsi, B.P. (1969). The multiple sample up-and-down method in bioassay. J. Amer. Statist. Assoc., 64, 147-162.
- (11) Noordwijk, A.J. van and J. van Noordwijk. (1988). An accurate method for estimating an approximate lethal dose with few animals tested with a Monte Carlo procedure. Arch. Toxicol., 61, 333-343.

- (12) SAS Institute Inc. (1990). SAS/STAT® User's Guide. Version 6, Fourth Ed. or later. Cary, NC, USA.
- (13) BMDP Statistics Software, Inc. (1990). BMDP Statistical Software Manual. W.J. Dixon, Chief Ed. 1990 rev. or later. University of California Press, Berkeley, CA, USA.
- (14) Lotus Development Corporation. (1999). Lotus® 1-2-3. Version 9.5, Millennium Edition. Cambridge, MA, USA.
- (15) Microsoft Corporation. (1985-1997). Microsoft® Excel. Version 5.0 or later. Seattle, WA, USA.

APPENDIX I

DEFINITIONS

Acute oral toxicity is the adverse effects occurring within a short time of oral administration of a single dose of a substance or multiple doses given within 24 hours.

Delayed death means that an animal does not die or appear moribund within 24 hours but dies later during the 14-day observation period.

Dosage is a general term comprising the dose, its frequency and the duration of dosing.

Dose is the amount of test substance administered. Dose is expressed as weight (g, mg) or as weight of test substance per unit weight of test animal (e.g. mg/kg).

LD50 (median lethal dose), oral, is a statistically derived single dose of a substance that can be expected to cause death in 50 per cent of animals when administered by the oral route. The LD50 value is expressed in terms of weight of test substance per unit weight of test animal (mg/kg).

Moribund status of an animal is the result of the toxic properties of a test substance where death is anticipated. For making decisions as to the next step in this test, animals killed for humane reasons are considered in the same way as animals that died.

Nominal sample size refers to the total number of tested animals reduced by one less than the number of like responses at the beginning of the series, or by the number of tested animals up to but not including the pair that creates the first reversal. For example, for a series as follows: OOOXXOXO, we have the total number of tested animals (or sample size in the conventional sense) as 8 and the nominal sample size as 6. It is important to note whether a count in a particular part of the guideline refers to the nominal sample size or to the total number tested. For example, the maximum actual number tested is 15. When testing is stopped based on that basis, the nominal sample size will be less than or equal to 15. Members of the nominal sample start with the animal numbered (r-1) (see reversal below).

Probit is an abbreviation for the term “probability integral transformation” and a probit dose-response model permits a standard normal distribution of expected responses (i.e., one centered to its mean and scaled to its standard deviation, *sigma*) to doses (typically in a logarithmic scale) to be analyzed as if it were a straight line with slope the reciprocal of *sigma*. A standard normal lethality distribution is symmetric; hence, its mean is also its true LD50 or median response.

Reversal is a situation where non-response is observed at some dose, and a response is observed at the next dose tested, or vice versa (i.e., response followed by non-response). Thus, a reversal is created by a pair of responses. The first such pair occurs at animals numbered r-1 and r.

Sigma is the standard deviation of a log normal curve describing the range of tolerances of test subjects to the chemical. *Sigma* provides an estimate of the variation among test animals in response to doses throughout the dose-response curve.

Slope (of the dose response curve) is the value that describes the angle at which the dose response curve rises from the dose axis. This value is the reciprocal of sigma.

APPENDIX II

DOSING PROCEDURE

Dose Sequence for Primary or Single-Sequence Test

1. For each run, animals are dosed, one at a time, at 48-hour intervals. The first animal receives a dose a step below the level of the best estimate of the LD50. This selection reflects an adjustment for a tendency to upward bias in the final estimate (see paragraph 5); as the test progresses, dosing will adjust for the overall pattern of outcomes. If the animal survives, the dose for the next animal is increased to a factor of 3.2 times the original dose; if it dies, the dose for the next animal is decreased by a similar dose progression. (Note: 3.2 is the default factor. Paragraph 3 below provides further guidance for choice of dose spacing factor). Each animal should be observed carefully for 48 hours (unless the animal dies) before making a decision on whether and how much to dose the next animal. That decision is based on the survival pattern of all the animals up to that time.

2. A combination of stopping criteria is used to keep the number of animals low while adjusting the dosing pattern to reduce the effect of a poor starting value. In any event, the test uses no more than 15 animals. Reaching one of the boundary doses and “staying there” for three animals stops the test. Unless this happens, the minimum number tested starting with the first reversal (called the nominal sample size) is 6. Testing stops at this point if and only if every response has been followed by a nonresponse or vice versa. (This outcome can be symbolized by ...XOXOXO or ...OXOXOX where X denotes dies within 48 hours, O denotes survives, and ... indicates a possible run of Xs or Os, respectively, preceding the example.) This type of outcome suggests the LD50 is very likely to be between the two particular test doses and that there is low variability in response sensitivity (e.g., a steep slope for an assumed probit dose-response model), a situation favorable for accurate results based on this guideline. Counting which contributes to the stopping decision is carried out from the first reversal to adjust for cases where there is an initial run of nonresponses or only responses, which tends to be associated with a poor starting dose. If there have been fewer than 5 reversals by this nominal sample size of 6, there is somewhat higher probability that more animals will be needed to achieve an accurate estimate. Possible problems include a relatively flat dose response, a starting value distant from the true LD50, an apparent adverse response not actually related to exposure to the test substance, or some combination of these factors. Therefore, in this case testing continues until it satisfies a criterion based on how likely it was to see the observed pattern, or the maximum allowable number of animals is reached.

3. Dose spacing is most successful if it can be related to the slope of dose response. At the outset, if feasible, a slope of the dose response should be estimated based on all information available to the toxicologist including structure activity relationships. The dose progression factor should be chosen to be the *sigma* or antilog of 1/(the estimated slope of the dose response curve). When there is no information on the substance to be tested, a dose progression factor of 3.2 is used.

4. Once the starting dose and dose spacing are decided, the toxicologist should list all possible doses including the upper (usually 2000 or 5000 mg/kg) and lower bounds. Doses that are close to the upper and lower bounds should be removed from the progression. Setting of lower bounds may need to include consideration of the ability to accurately dilute the test material).
5. The stepped nature of the TG 425 design provides for the first few doses to function as a self-adjusting sequence. Because of the tendency for positive bias, in the event that nothing is known about the substance, a starting dose of 175 mg/kg is recommended. If the default procedure is to be used for the primary test, dosing will be initiated at 175 mg/kg and doses will be spaced by a factor of 0.5 (\log_{10} dose). The doses to be used are 1, 5.5, 17.5, 55, 175, 550, 1750, 2000, or, for specific regulatory needs, 5000 instead of 2000.
6. Only the doses in the predetermined dose progression (either one analytically based or the default progression) should be used. This avoids changing the dose progression if either the upper or lower limit is reached during the study. If there is no reversal before reaching either the upper or lower bounds, no more than three animals should be dosed at these limiting doses (see stopping criterion (1) in paragraph 20).

Setting Starting Doses for Supplemental Multi-Sequence Procedure

7. In order to maximize information on the dose response curve, the starting doses of each sequence should be staggered in such a way that the doses tested in one sequence are between the doses of neighboring sequences. The factor 3.2 comes from the fact that this value forces alternating doses in the full list of possible doses to be separated by approximately one order of magnitude, i.e., a 10-fold difference. For example, the dose list 1, 3.2, 10, 32, 100... is one where every other dose is separated by a 10-fold increment. Furthermore, the same list, on the base 10 log-scale is 0.0, 0.5, 1.0, 1.5, 2.0... which illustrates the fact that a constant multiplicative factor separating doses on the mg/kg dose scale translates to an additive equal spacing on the base 10 log scale. It also exhibits the fact that $\log_{10}(3.2) = 0.5$, i.e., one-half of one order of magnitude.
8. By working on the log-scale, staggering doses is straightforward. On that scale, one need only partition the log-scale dosing increment into the number of staggered start doses needed. For example, $0.5/5 = 0.1$, so that starting doses for five separate sequences could be 1.0, 1.1, 1.2, 1.3, 1.4 on the log-scale, which translates to 10.0, 12.6, 15.8, 20.0, 25.1. The next dose in this list of starting doses, 1.5 (or 31.6), is the next dose in the testing sequence that starts at 1.0 (or 10.0). It is also worth noting that the factor that separates each starting dose on the actual dose scale, 1.26, is the fifth-root of 3.2.
9. The specific steps to be followed are:
 1. Select a dose about which one wishes to stagger doses.
 2. Convert the dose in (1) to log-scale, and calculate the \log_{10} of the dosing increment.
 3. Divide the log of the dosing increment by the number of sequences to be use.
 4. Add or subtract the dosing increment to the dose in (1), repeatedly until the correct number of starting doses is created.

5. Convert the log doses back to the original scale.

10. As a second example, (1) Suppose we want to stagger four starting doses around a dose of 120, and the dosing increment is 3.2. (2) The log starting value is $\log_{10}(120) = 2.079$, and $\log_{10}(3.2) = 0.5$. For step (3), $0.5/4 = .125$. (4) Since there are an even number of starts, we will put 2 starts below 120, and one above. The starts below 120 are $2.079 - 0.125 = 1.954$, $1.954 - 0.125 = 1.829$. The start above 120 is $2.079 + 0.125 = 2.204$, or together, 1.829, 1.954, 2.079, 2.204. (5) Finally, converting the original dose scale, these starts are 67, 90-, 120 160.

Limit Test

11. The Limit Test is a sequential test that may use up to 5 animals. A test dose of up to 2000 (and exceptionally 5000) mg/kg may be used.

12. Dose one animal at the test dose. If the animal dies, conduct the primary test to determine the LD50. If the animal survives, dose two additional animals. If both animals survive, the LD50 is greater than the limit dose and the test is terminated. If one or both animals die, then dose an additional two animals, one at a time. The results are evaluated as follows (S=survival, D=death).

13. The LD50 is less than test dose (2000 mg/kg or 5000 mg/kg) when three or more animals die.

S DS DD
S SD DD
S DD DX
S DD SD
S DD DX

14. The LD50 is greater than the test dose (2000 mg/kg or 5000 mg/kg) when three or more animals survive.

S DS DS
S DS SX (X can be S or D, the dosing of 5th animal is not necessary)
S SD DS
S SD SX (X can be S or D, the dosing of 5th animal is not necessary)
S DD SS

APPENDIX III

Computations for the Likelihood-Ratio Stopping Rule

As described in Guideline paragraph 20, a likelihood-ratio stopping rule is evaluated after testing each animal, starting with the fourth tested following the reversal. Three "measures of test progress" are calculated. Technically, these measures of progress are likelihoods, as recommended for the maximum-likelihood estimation of the LD50. The procedure is closely related to calculation of a confidence interval by a likelihood-based procedure.

The basis of the procedure is that when enough data have been collected, a point estimate of the LD50 should be more strongly supported than values above and below the point estimate, where statistical support is quantified using likelihood. Therefore three likelihood values are calculated, a likelihood for an LD50 point estimate, a likelihood for a value below the point estimate, and a likelihood for a value above the point estimate. Specifically, the low value is taken to be the point estimate divided by 2.5 and the high value is taken to be the point estimate multiplied by 2.5.

The likelihood values are compared by calculating ratios of likelihoods, and then determining whether the likelihood ratios (LR) exceed a critical value. Testing stops when the ratio of the likelihood for the point estimate exceeds each of the other likelihoods by a factor of 2.5, which is taken to indicate relatively strong statistical support for the point estimate. Therefore two likelihood ratios (LRs) are calculated, a ratio of likelihoods for the point estimate and the point estimate divided by 2.5, and a ratio for the point estimate and the estimate times 2.5. The values of 2.5 here have been shown using simulations to yield a useful stopping rule.

The calculations are easily performed in any spreadsheet with normal probability functions. The calculations are illustrated in the following table, which is structured to promote spreadsheet implementation. The computation steps are illustrated using an example where the upper boundary dose is 5000 mg/kg, but the computational steps are identical when the upper boundary dose is 2000 mg/kg. Empty spreadsheets preprogrammed with the necessary formulas are available for direct downloading on the OECD and EPA websites.

Hypothetical example using upper boundary 5000 mg/kg (Table 1)

In the hypothetical example utilizing an upper boundary dose of 5000 mg/kg, the LR stopping criterion was met after nine animals had been tested. The first "reversal" occurred with the 3rd animal tested. The stopping criterion is checked when four animals have been tested following the reversal. In this example, the fourth animal tested following the reversal is the seventh animal actually tested. Therefore, for this example, the data would have been entered into the spreadsheet only after the seventh animal had been tested. Subsequently, the stopping criterion would have been checked after testing the seventh animal, the eighth animal, and the ninth. The stopping criterion is satisfied after the ninth animal is tested.

A. Enter the dose-response information.

After each animal is tested, the results are entered at the end of the matrix in Columns 1-4.

Column 1. Steps are numbered 1-15. A maximum of 15 animals may be tested.

Column 2. Enter the dose received by the i^{th} animal.

Column 3. Indicate whether the animal responded (we use an X) or did not respond (we use an O).

The results should be entered in the same order as animals are tested.

B. The nominal and actual sample sizes.

The nominal sample consists of the two animals that represent the reversal (here the second and third), plus all animals tested subsequently. Here, we use Column 4 to indicate whether or not a given animal is included in the nominal sample.

- Enter the nominal sample size (nominal n) in Row 16. This is the number of animals in the nominal sample. In the example, nominal n is 8.
- Enter the actual number tested in Row 17.

C. Rough estimate of the LD50.

As a rough estimate of the LD50 from which to gauge progress, we use the geometric mean of doses for the animals in the nominal sample. In the table, this is called the “dose-averaging estimator.” We restrict this average to the nominal sample in order to allow for a poor choice of initial test dose, which could generate either an initial string of non-responses or an initial string of non-responses. (However, we will use the results for all animals in the likelihood calculations below.) Recall that the geometric mean of n numbers is the product of the n numbers, raised to a power of $1/n$.

- Enter the dose-averaging estimate in Row 18. In the example, the value in Row 18 is equal to $(320 (1000 (\dots (1000)^{1/8} = 754$.
- Enter in Row 19 the logarithm (base 10) of the value in Row 18. The value in Row 19 is $\log_{10} 754 = 2.9$.

A more refined procedure could use the maximum-likelihood estimate of the LD50. The dose-averaging estimator is used to simplify the calculations.

D. Likelihood for the crude LD50 estimate.

“Likelihood” is a statistical measure of how strongly the data support an estimate of the LD50 or other parameter. Ratios of likelihood values can be used to compare how well the data support different estimates of the LD50.

In Column 7 we calculate the likelihood for the estimate of the LD50 that was calculated at Step C. The likelihood (Row 21) is the product of likelihood contributions for individual animals. The likelihood contribution for the i^{th} animal is denoted L_i . (In our implementation, we use the algebraically equivalent approach of summing the logarithms of the L_i values, then taking the antilog of the sum.)

Column 6. Enter the estimate of the probability of response at dose d_i , denoted P_i . P_i is calculated from a dose-response curve. Note that the parameters of the probit dose-response curve are the slope and the LD50, so values are needed for each of those parameters. For the LD50 we use the dose-averaging estimate from Row 18. For the slope we use the default value of 2. The following steps may be used to calculate the response probability P_i .

1. Calculate the base-10 log of dose d_i (Column 5).
2. For each animal calculate the z-score, denoted Z_i (not shown in the table), using the formulae

$$\begin{aligned}\sigma &= 1 / \text{slope}, \\ Z_i &= (\log_{10}(d_i) - \log_{10}(\text{LD50})) / \sigma\end{aligned}$$

For example, for the first animal (Row 1), we have

$$\begin{aligned}\sigma &= 1 / 2 \\ Z_1 &= (2.000 - 2.878) / 0.500 = -1.756\end{aligned}$$

3. For the i^{th} dose the estimated response probability is

$$P_i = F(Z_i)$$

where F denotes the cumulative distribution function for the standard normal distribution (i.e., the normal distribution with mean 0 and variance 1).

For example (Row 1), we have

$$P_1 = F(-1.756) = 0.0396$$

The function F (or something very close) is ordinarily what is given for the normal distribution in statistical tables, but the function is also widely available as a spreadsheet function. It is available under different names, for example the @NORMAL function of Lotus 1-2-3 (14) and the @NORMDIST function in Excel (15). To confirm that you have used correctly the function available in your software, you may wish to verify familiar values such as $F(1.96) = 0.975$ or $F(1.64) = 0.95$.

Column 7. Calculate the natural log of the likelihood contribution ($\ln(L_i)$). L_i is simply the probability of the response that actually was observed for the i^{th} animal:

$$\begin{aligned}\text{responding animals: } \ln(L_i) &= \ln(P_i) \\ \text{non-responding animals: } \ln(L_i) &= \ln(1 - P_i)\end{aligned}$$

Note that here we have used the natural logarithm (ln), whereas elsewhere we use the base-10 (common) logarithm. These choices are what are ordinarily expected in a given context.

The steps above are performed for each animal. Finally:

Row 20: Sum the log-likelihood contributions in Column 7.

Row 21: Calculate the likelihood by applying the exp function applied to the log-likelihood value in Row 20. In the example, $\exp(-3.385) = e^{-3.385} = 0.0338$.

E. Calculate likelihoods for two dose values above and below the crude estimate.

If the data permit a precise estimate, then the likelihood should be high for a reasonable estimate of the LD50, relative to likelihoods for values distant from our estimate. We compare the likelihood for the dose-averaging estimate (754, Row 18) to values differing by a factor of 2.5 from that value (i.e., to 754×2.5 and $754/2.5$). The calculations (displayed in Columns 8-11) are similar to those described above, except that the values 301.7 ($=754/2.5$) and 1986 ($=754 \times 2.5$) have been used for the LD50, instead of 754. The likelihoods and log-likelihoods are displayed in Rows 20-21.

F. Calculate likelihood ratios.

The three likelihood values (Row 21) are used to calculate two likelihood ratios (Row 22). A likelihood ratio is used to compare the statistical support for the estimate of 754 to the support for each of the other values, 301.7 and 1985.9. The two likelihood ratios are therefore:

$$\begin{aligned}\text{LR1} &= [\text{likelihood of 754}] / [\text{likelihood of 301.7}] \\ &= 0.0338 / 0.0082 \\ &= 4.10\end{aligned}$$

and

$$\begin{aligned}\text{LR2} &= [\text{likelihood of 754}] / [\text{likelihood of 1985.9}] \\ &= 0.0338 / 0.0097 \\ &= 3.49\end{aligned}$$

G. Determine if the likelihood ratios exceed the critical value.

High likelihood ratios are taken to indicate relatively high support for the point estimate of the LD50. Both of the likelihood ratios calculated in Step F (4.10 and 3.49) exceed the critical likelihood ratio that we use, which is 2.5. Therefore the LR stopping criterion is satisfied and testing stops.

TABLE
1

	1	2	3	4	5	6	7	8	9	10	11
	Step <i>I</i>	Dose	(X) response (O) non-resp.	Included in nominal <i>n</i>	log10 Dose <i>d_i</i>	LD50 =	794.1	LD50 =	301.7	LD50 =	1885.9
						Prob. of response	ln(<i>L_i</i>)	Prob. of response	ln(<i>L_i</i>)	Prob. of response	ln(<i>L_i</i>)
1	1	100	O	NO	2.00	0.0396	-0.0404	0.1687	-0.1848	0.0054	-0.0054
2	2	320	O	YES	2.50	0.2282	-0.2590	0.5203	-0.7347	0.0617	-0.0637
3	3	1000	X	YES	3.00	0.5967	-0.5163	0.8510	-0.1613	0.2908	-1.2351
4	4	320	O	YES	2.50	0.2282	-0.2590	0.5203	-0.7347	0.0617	-0.0637
5	5	1000	X	YES	3.00	0.5967	-0.5163	0.8510	-0.1613	0.2908	-1.2351
6	6	320	O	YES	2.50	0.2282	-0.2590	0.5203	-0.7347	0.0617	-0.0637
7	7	1000	O	YES	3.00	0.5967	-0.9081	0.8510	-1.9038	0.2908	-0.3436
8	8	3200	X	YES	3.70	0.8953	-0.1106	0.9799	-0.0203	0.6770	-0.3901
9	9	1000	X	YES	3.00	0.5967	-0.5163	0.8510	-0.1613	0.2908	-1.2351
10	10			-	-	-	-	-	-	-	-
11	11			-	-	-	-	-	-	-	-
12	12			-	-	-	-	-	-	-	-
13	13			-	-	-	-	-	-	-	-
14	14			-	-	-	-	-	-	-	-
15	15			-	-	-	-	-	-	-	-
16	Nominal Sample size =				8						
17	Actual number tested =				9						
18	Dose-averaging estimator				754.35						
19	log10 =				2.878						
20	log-likelihood sums:										
21	likelihoods:										
22	likelihood ratios:										
23	Individual ratios exceed critical=				2.5			TRUE		TRUE	
24	Both ratios exceed critical value?							TRUE			

APPENDIX IV

Alternate Supplemental Procedure

The design for slope estimation involves multiple stages of testing. The first stage is execution of the Primary Procedure. Subsequent stages involve concurrent up-and-down testing sequences with nominal sample size 2, with (at each stage) some sequences initiated at a relatively low dose and others at a higher dose, compared to the LD50. This design is considered to provide adequate precision for estimation of the slope in most situations. (It is thought that the precision required will not usually exceed the precision provided by the design.) If there are situations where the required precision can be stated precisely, it may be possible to reduce the number of animals tested by terminating the study, when the data collected up to a given point permit an estimate with the precision required.

The design has 5 stages. At Stages 2 and following, all testing sequences have nominal sample size of two, i.e., the sequence terminates when a reversal is observed.

Stage 1: Execute the primary procedure, with the guideline stopping criteria.

Stage 2: Execute two up-and-down testing sequences, each with successive test doses spaced by 2 log units (a progression factor of 100). One sequence is started at a low dose relative to the LD50 and the other at a high dose relative to the LD50.

Stage 3: Execute 2 sequences with doses spaced by 0.5 log unit (a factor of approximately 3.2), one starting at a low dose and one starting at a high dose, relative to the LD50.

Stage 4: Execute 2 sequences with doses spaced by 0.25 log units, one starting at a low dose and one at a high dose, relative to the LD50.

Stage 5: Execute 3 sequences with doses spaced by 0.125 log units, 2 starting at a low dose and one at a high dose, relative to the LD50.

The following procedure is to be used for selecting initial test doses, for up-and-down sequences at Stage 2 and following. Where the intent is for the sequence to be initiated at a low dose relative to the LD50, the initial test dose equals the highest dose tested, such that an adverse affect has not been observed at that dose, or at any lower doses tested, considering the results of all completed stages of the study. Where the intent is for the sequence to be initiated above the LD50, the initial test dose is chosen to equal the lowest test dose that is associated with 100% response in all tests of that dose, as well as at all higher tested doses. In cases where the lowest dose tested is associated with an adverse effect for one or more animal, the initial test dose is chosen to equal that dose, divided by the progression factor for the current stage. In cases where the highest dose tested is associated with no adverse effects, the initial test dose is chosen to equal that dose, multiplied by the progression factor for the current stage.

Where the range of test doses is restricted (e.g., if the test doses may not exceed 2000 units or may not exceed 5000 units), and the application of these criteria would result in a dose beyond a bound of the range, the dose is chosen to equal the corresponding bounding dose (e.g., chosen equal to 2000 units or 5000 units). Whenever a bounding dose is tested, the next dose to be tested (in the same sequence) may equal the same bounding dose, or may be chosen strictly within the dose range, based on precisely the same criteria as for the Primary Procedure. As for

the Primary Procedure, a single up-and-down testing sequence is stopped if three successive test doses equal a bounding dose, with no responses (when the dose is an upper bound dose) or with three responses (for a lower bound dose).

The number of animals that can be tested is restricted as follows. Upon completion of a given stage, testing stops if the number tested (in that stage and previous stages) equals or exceeds 40. The minimum number, based on the minimum nominal sample size for each sequence, is 24 ($=6 + 2*2 + 2*2 + 2*2 + 3*2$). In practice, it is believed that the numbers tested will usually not exceed 40.

After all stages of the test are completed, results of all stages are combined in a single probit analysis. The statistics reported are to include confidence intervals for the slope and LD50, as well as point estimates for those parameters, where available, calculated using standard procedures of probit analysis.

